The Hypercube Structure of the Genetic Code Explains Conservative and Non–Conservative Aminoacid Substitutions *in vivo* and *in vitro*

Miguel A. Jiménez–Montaño*

Departamento de Física y Matemáticas, Universidad de las Américas,
Puebla, Sta. Catarina Mártir, 72820 Cholula, Puebla, México

Carlos R. de la Mora–Basáñez[+]

Dirección General de Investigaciones, Universidad Veracruzana,
Xalapa, Ver. 91000, México

Thorsten Pöschel**

Arbeitsgruppe Nichtlineare Dynamik, Universität Potsdam, Am Neuen Palais, D-14415 Potsdam, Germany, and
The James Franck Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637

(February 1, 2008)

A representation of the genetic code as a six–dimensional Boolean hypercube is described. This structure is the result of the hierarchical order of the interaction energies of the bases in codon–anticodon recognition. In this paper it is applied to study molecular evolution *in vivo* and *in vitro*. In the first case we compared aligned positions in homologous protein sequences and found two different behaviors: a) There are sites in which the different amino acids may be explained by one or two "attractor nodes" (coding for the dominating amino acid(s)) and their one–bit neighbors in the codon hypercube, and b) There are sites in which the amino acids correspond to codons located in closed paths in the hypercube. In the second case we studied the "Sexual PCR"[1] experiment described by Stemmer [1] and found that the success of this combination of usual PCR and recombination is in part due to the Gray code structure of the genetic code.

PACS numbers:

I. INTRODUCTION

The genetic code is the biochemical system for gene expression. It deals with the translation, or decoding, of information contained in the primary structure of DNA and RNA molecules into protein sequences. Therefore the genetic code is both, a physico–chemical and a communication system. Physically, molecular recognition depends on the degree of complementarity between the interacting molecular surfaces (by means of weak interactions); informationally, a prerequisite to define a code is the concept of distinguishability. It is the physical indistinguishability of some codon–anticodon interaction energies that makes the codons synonymous, and the code degenerate and redundant [2].

In natural languages [3] as well as in the genetic code the total redundancy is due to a hierarchy of constraints acting one upon another. The specific way in which the code departs from randomness is, by definition, its structure. It is assumed that this structure is the result of the hierarchical order of the interaction energies of the bases in codon–anticodon recognition. The hypercube structure of the genetic code as currently introduced [4] will be described and its implications for molecular evolution and test–tube evolution experiments will be discussed. As we shall see the genetic code may be represented by a six–dimensional boolean hypercube in which the codons (actually the code–words; see below) occupy the vertices (nodes) in such a way that all kinship[2] neighborhoods are correctly represented. This approach is a particular application to binary sequences of length six of the general concept of sequence–space, first introduced in coding theory by Hamming [5].

A code–word is next to six nodes representing codons differing in a single property. Thus the hypercube simultaneously represents the whole set of codons and keeps track of which codons are one–bit neighbors of each other. Different hyperplanes correspond to the four stages of the evolution of the code according to the Co–evolution Theory [6–8]. Transitions within three of the

---

[1]Polymerase Chain Reaction plus DNA shuffling
[2]The term kinship means the relationship between members of the same family.

"columns" (four–dimensional cubes), consisting of the codon classes NGN, NAN, NCN, and NUN, lead to silent and conservative amino acid substitutions; while transitions in the same hyperplane (four–dimensional subspace belonging to any of the codon classes ANN, CNN, GNN or UNN) lead to non–conservative substitutions as frequently found in proteins. The proposed structure demonstrates that in the genetic code there is a good balance between conservatism and innovation. To illustrate these results several examples of the non–conservative variable positions of homologous proteins are discussed. Two different behaviors were found:

i There are sites in which the different amino acids may be explained by one or two "attractor nodes" (coding for the dominating amino acid(s)) and their one–bit neighbors in the codon hypercube, and

ii There are sites in which the amino acids correspond to codons located in closed paths in the hypercube.

Very recently the rapid evolution of a protein *in vitro* by DNA shuffling has been accomplished by Stemmer [1].

This experiment, called by Smith "Sexual PCR", was further discussed in [9]. Smith recalls that Stemmer investigated the *beta*–lactamase gene TEM–1 which has a very low activity against the antibiotic cefoxtamine[3] After three cycles of mutagenesis, recombination and selection he found the minimum inhibitory concentration to be $16,000$ times higher than that of the original clone.

It will be shown that, without exception, the amino–acid replacements in TEM–1 mutants selected for high resistance to cefotaxime may be accounted by one bit changes of the corresponding codons. This shows that the structure of the code permits a very significant change in function of the coded protein by means of one–bit changes of some of the codons, provided that these mutations are integrated in a single polynucleotide by recombination.

II. CODON–ANTICODON INTERACTION

The four bases occurring in DNA (RNA) macromolecules define the corresponding alphabet X : $\{A, C, G, T\}$ or X : $\{A, C, G, U\}$. Each base is completely specified by two independent dichotomic categorizations (Fig. 1):

i according to its chemical type $\mathcal{C}$: $\{R, Y\}$, where R : (A, G) are purines and Y : (C, U) are pyrimidines and

ii according to H–bonding, $\mathcal{H}$: $\{W, S\}$, where W : (A, U) are weak and S : (C, G) are strong bases.
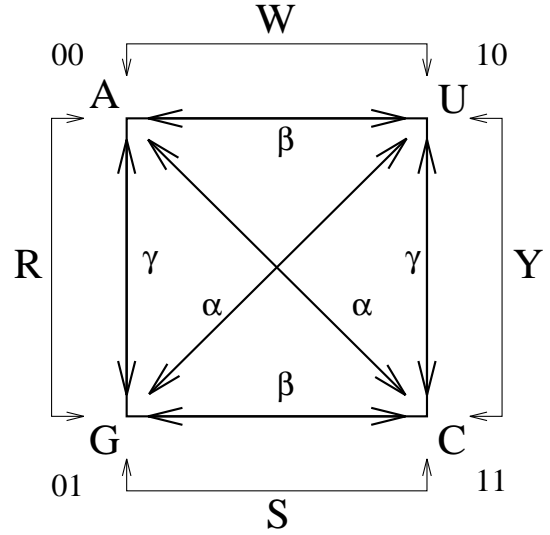

FIG. 1. Categorizations of the bases. The categorizations of the bases according to (i): chemical type $\mathcal{C}$: $\{R, Y\}$ where R : (A, G) are purines and Y : (C, U) are pyrimidines, and (ii) according to H–bonding, $\mathcal{H}$: $\{W, S\}$, where W : (A, U) are weak and S : (C, G) strong bases. The third possible partition into imino/keto bases is not independent from the former ones and is irrelevant for the codon–anticodon interaction. The binary representation of the bases is also shown. The first bit is the chemical type and the second one the H–bonding character. α, β and γ are the transformations of the bases which form a Klein–4 group [8,11].

The third possible partition into imino/keto bases is not independent from the former ones. Denoting by $\mathcal{C}_i$ the chemical type and by $\mathcal{H}_i$ the H–bond category of the base B_i at position i of a codon our basic assumption says that the codon–anticodon interaction energy obeys the following hierarchical order:

$$\mathcal{C}_2 > \mathcal{H}_2 > \mathcal{C}_1 > \mathcal{H}_1 > \mathcal{C}_3 > \mathcal{H}_3 \ .$$

This means, that the most important characteristic determining the codon–anticodon interaction is the chemical type of the base in the second position; the next most important characteristic is whether there is a weak or strong base in this position; then the chemical type of the first base and so on.

The above assumption goes beyond the early qualitative view that the optimization between stability and rate, that is always found for enzyme–substrate inter-

---

[3]The minimum inhibitory concentration for Escherichia Coli bacteria carrying TEM–1–bearing plasmid is only $20 \ ng \ ml^{-1}$.

actions, also applies to the codon–anticodon interaction [10]. Besides, several authors have suggested that three bases are needed for effectively binding the adapter to the messenger. From this it maybe inferred that codon's size determines a range of codon–anticodon overall interaction strength within which recognition can occur. Genetic translation rate is limited, among other things, by codon–anticodon recognition which, in turn, depends on base–pair lifetimes in a given structural situation. These life–times are influenced by the nature of the pairs: they are shorter for $A - T$ than for $G - C$ pairs [11].

The bases are represented by the nodes of a 2–cube (Fig. 1). The first attribute is the chemical character and the second one is the hydrogen–bond character. Extending this association to base triplets, each codon is associated in a unique way with a codeword consisting of six attribute values (see Table 1).

0	0	0	0	1	1	A	A	C	N
0	0	0	0	1	0	A	A	U	N
0	0	0	0	0	0	A	A	A	K
0	0	0	0	0	1	A	A	G	K
1	0	0	0	0	1	U	A	G	t
1	0	0	0	0	0	U	A	A	t
1	0	0	0	1	0	U	A	U	Y
1	0	0	0	1	1	U	A	C	Y
1	1	0	0	1	1	C	A	C	H
1	1	0	0	1	0	C	A	U	H
1	1	0	0	0	0	C	A	A	Q
1	1	0	0	0	1	C	A	G	Q
0	1	0	0	0	1	G	A	G	E
0	1	0	0	0	0	G	A	A	E
0	1	0	0	1	0	G	A	U	D
0	1	0	0	1	1	G	A	C	D
0	1	1	0	1	1	G	U	C	V
0	1	1	0	1	0	G	U	U	V
0	1	1	0	0	0	G	U	A	V
0	1	1	0	0	1	G	U	G	V
1	1	1	0	0	1	C	U	G	L
1	1	1	0	0	0	C	U	A	L
1	1	1	0	1	0	C	U	U	L
1	1	1	0	1	1	C	U	C	L
1	0	1	0	1	1	U	U	C	F
1	0	1	0	1	0	U	U	U	F
1	0	1	0	0	0	U	U	A	L
1	0	1	0	0	1	U	U	C	L
0	0	1	0	0	1	A	U	G	M
0	0	1	0	0	0	A	U	A	I
0	0	1	0	1	0	A	U	U	I
0	0	1	0	1	1	A	U	C	I
0	0	1	1	1	1	A	C	C	T
0	0	1	1	1	0	A	C	U	T
0	0	1	1	0	0	A	C	A	T
0	0	1	1	0	1	A	C	G	T
1	0	1	1	0	1	U	C	G	S
1	0	1	1	0	0	U	C	A	S
1	0	1	1	1	0	U	C	U	S

1	0	1	1	1	1	U	C	C	S
1	1	1	1	1	1	C	C	C	P
1	1	1	1	1	0	C	C	U	P
1	1	1	1	0	0	C	C	A	P
1	1	1	1	0	1	C	C	G	P
0	1	1	1	0	1	G	C	G	A
0	1	1	1	0	0	G	C	A	A
0	1	1	1	1	0	G	C	U	A
0	1	1	1	1	1	G	C	C	A
0	1	0	1	1	1	G	G	C	G
0	1	0	1	1	0	G	G	U	G
0	1	0	1	0	0	G	G	A	G
0	1	0	1	0	1	G	G	G	G
1	1	0	1	0	1	C	G	G	R
1	1	0	1	0	0	C	G	A	R
1	1	0	1	1	0	C	G	U	R
1	1	0	1	1	1	C	G	C	R
1	0	0	1	1	1	U	G	C	C
1	0	0	1	1	0	U	G	U	C
1	0	0	1	0	0	U	G	A	t
1	0	0	1	0	1	U	G	C	W
0	0	0	1	0	1	A	G	G	R
0	0	0	1	0	0	A	G	A	R
0	0	0	1	1	0	A	G	U	S
0	0	0	1	1	1	A	G	C	S

TABLE I. Gray code representation of the genetic code. In the first and fourth blocks the six–dimensional vectors (code–words) are shown. In the second and fifth blocks appear the corresponding codons. Finally, in the third and sixth columns the amino acids in single letter notation. The first two digits correspond to the first base, the following two to the second base and the last two to the last base, according to the binary codification of the bases of Fig. 1.

In some of the hypercube directions single feature codon changes (one–bit code–word changes) produce synonymous or conservative amino acid substitutions in the corresponding protein (when the transitions occur in three of the 4–cubes displayed as "columns" in Figs. 2 and 4); while in other directions lead to context dependent replacements which, in general, conserve only certain physical properties. However, if these properties are the only relevant ones in the given context, the substitution has little effect on the protein structure as well. These low–constraint sites facilitate evolution because they allow the transit between hypercube columns belonging to amino acids with very different physico–chemical properties (e.g. hydrophobic and hydrophilic amino acids, respectively).
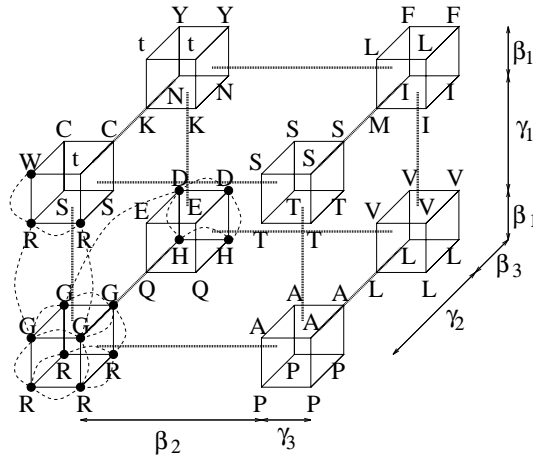
FIG. 2. The six–dimensional hypercube. Each node is labeled with the corresponding amino acid in the single letter notation or terminator symbol. The fat short dashed lines represent a complex connection between two (three–dimensional) cubes. Such a line represents 8 edges each, connecting the corresponding nodes of two neighbored three–dimensional cubes (see fig. 3). The cluster of amino acids of the first example discussed in the text is displayed by fat points at the corresponding nodes and dashed thin curved lines for the edges.
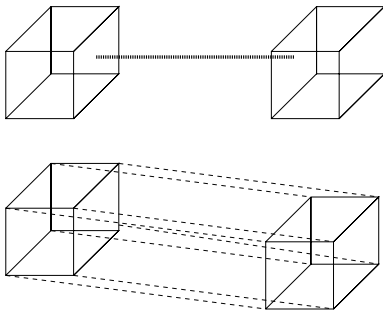

FIG. 3. Each of the fat short dashed lines represent 8 edges, connecting the corresponding nodes of two three–dimensional cubes. The figure shows a four–dimensional cube using the symbolic fat drawn link (top) and the same cube using standard representation.
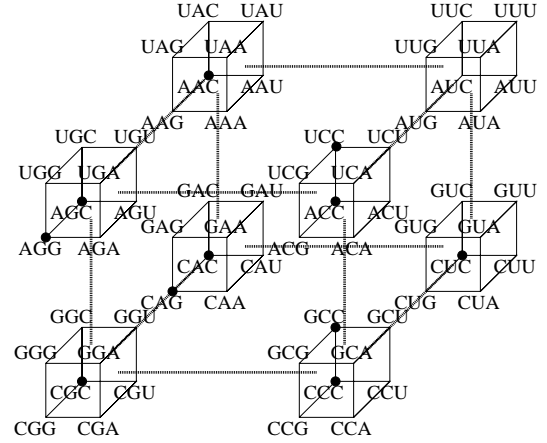

FIG. 4. The hypercube representation of the genetic code. Each node represents a code–word (six–dimensional vector) of attribute values. However, for clarity of interpretation, the nodes are labeled with the corresponding codons (See Table 1 for the assignment of codons to vectors). The nodes and links mentioned in second example discussed in the text are shown. The edges connect: $AGG \leftrightarrow AGC$, $AGC \leftrightarrow ACC$, $AGC \leftrightarrow AAC$, $UCC \leftrightarrow ACC$, $ACC \leftrightarrow GCC$, $GCC \leftrightarrow CCC$, $CGC \leftrightarrow CAC$, $CAC \leftrightarrow CAG$, $CAC \leftrightarrow CUC$, $CAC \leftrightarrow GAC$

III. GRAY CODE STRUCTURE OF THE GENETIC CODE

An n–dimensional hypercube, denoted by Q_n, consists of 2^n nodes each addressed by a unique n–bit identification number. A link exists between two nodes of Q_n if and only if their node addresses differ in exactly one bit position. A link is said to be along dimension i if it connects two nodes which addresses differ to as the ith bit (where the least significant bit is referred to as the 0th bit). Q_6 is illustrated in Fig. 4. Two nodes in a hypercube are said to be adjacent if there is a link between them. The (Hamming) distance between any two cube nodes is the number of bits differing in their addresses. The number of transitions needed to reach a node from another node equals the distance between the two nodes. A d–dimensional sub-cube in Q_n involves 2^d nodes which addresses belong to a sequence of n symbols $\{0, 1, *\}$ in which exactly d of them are of the symbol $*$ (i.e. the don't care symbol which value can be 0 or 1).

The idea to propose a Gray Code representation of the Genetic Code goes back to Swanson [12] where this concept is explained in detail (see also [13]). However, a

great number of different Gray Codes can be associated to the Genetic Code depending on the order of importance of the bits in a code–word. In Table 1 our chosen Gray Code is displayed. It is constructed according to our main hypothesis

$$\mathcal{C}_2 > \mathcal{H}_2 > \mathcal{C}_1 > \mathcal{H}_1 > \mathcal{C}_3 > \mathcal{H}_3 \ .$$

For example, the first two lines of the table differ in the last bit corresponding to $\mathcal{H}_3$; which is the least significant bit; the second and the third lines differ in the next least significant bit, i.e. $\mathcal{C}_3$, and so forth.

IV. THE STRUCTURE OF CODON DOUBLETS

This section is more mathematical than the rest of the paper. It is not essential for the understanding of the rest of the paper.

In a pioneering paper Danckwerts and Neubert [14] discussed the symmetries of the sixteen $B_1 B_2$ codon doublets in terms of the Klein–4 group of base transformations. Here their result will be recast in a form of a decision–tree (Fig. 5) and their analysis will be extended to the $B_2 B_3$ doublets. They found the following structure for the set M of $B_1 B_2$ doublets:

Starting from Ac generate the set:

$$
\begin{aligned}
M_0 &= \{[(1,1) \cup (\alpha,1) \cup (\alpha,\beta) \cup (\alpha,\gamma)] \, AC\} \\
&= \{AC, CC, CG, CU\} \\
M_1 &= [(1,1) \cup (\beta,1)] \, M_0 \\
M_2 &= (\alpha,\alpha) \, M_1
\end{aligned}
$$

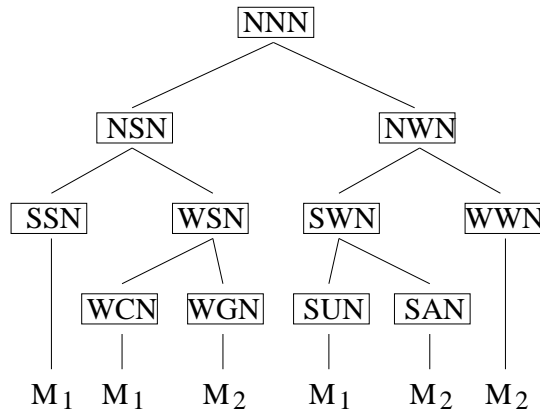The sets M_1 and M_2 consist of four–fold and less than four–fold degenerate doublets, respectively.

FIG. 5. Decision–tree of codon categories and redundancy distribution. The leaves are the sets of four–fold (M_1) and less than four–fold (M_2) degenerate $B_1 B_2$ doublets.

The set M can be expressed as:

$$M = [(1,1) \cup (b,1)] \, [(1,1) \cup (\alpha,\alpha)] \, M_0$$

Where the base exchange operators α, β, γ are defined in Fig. 1.

They showed that: "a) M_1 and M_2 are invariant by operating with $(\beta, 1)$ on B_1, but no operation on B_2 leaves M_1 or M_2 invariant. Thus B_2 carries more information than B_1 and B_2 is therefore more important for the stability of M_1 and M_2 than $B_1 \ldots A$ change of B_1 with respect to its hydrogen bond property does not change the resulting amino acids if all doublets of either M_1 or M_2 are affected.

Reversing supposition and conclusion, M_1 and M_2 may be defined as those doublet sets of 8 elements which are invariant under the $(\beta, 1)$–transformation. Then experience shows that M_1 and M_2 are fourfold and less than fourfold degenerate respectively."

Thus the third base degeneracy of a codon does not depend on the exact base B_1, but only on its H–bond property (weak or strong).

The above results can be simply visualized as a decision–tree (Fig. 5). It can be seen from this figure that the redundancy of a codon is determined only by the H–bond character of B_1 and B_2: SSN codons (with 6 H–bonds in $B_1 B_2$) belong to M_1 while WWN codons (with 4 H–bonds in $B_1 B_2$) belong to M_2. However, for codons WSN and SWN (with 5 H–bonds in $B_1 B_2$) it is not possible to decide unless one has more information about the second base: WCN and SUN belong to M_1 while WGN and SAN belong to M_2. In all cases at most three attributes are necessary to determine the redundancy of a codon up to this point. Of course the non–degenerate codons (UAG for Methionine and UGG for Tryptophan) will require the specification of the six attributes.

From the decision rules obtained from Fig. 5 it is clear that there are branches where the refinement procedure cannot continue (the branches which end in M_1) because no matter which base occupies the third codon position the degeneracy cannot be lifted. This imposes a limit to the maximum number of amino acids which can be incorporated to the code without recurring to a "frozen accident" hypothesis. Our proposal generalizes the "2–out–of–3" hypothesis of Lagerkvist [15] which refers only to codons in the SSN class.

The sixteen $B_1 B_2$ doublets can be represented as the vertices of a four–dimensional hypercube. Figure 6 shows that the sets M_1 and M_2 are located in compact regions. Notice that this figure differs from the one introduced by Bertman and Jungck [16] who considered as basic transformations α and β, instead of β and γ as we did. Since the operator α changes two bits we do not consider it as basic.
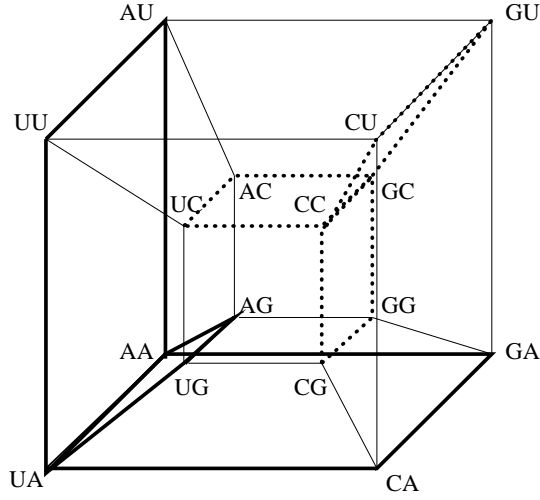
FIG. 6. The four–dimensional hypercube representation of the sets M_1 (dotted) and M_2 (fat).
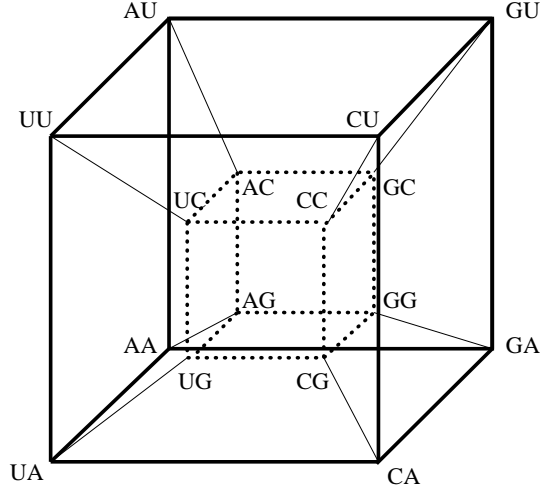

FIG. 7. The corresponding to fig. 6 hypercube of the sets M_1' and M_2' (see text). Notice that in both cases each set is located in a compact region.

Let's consider now the structure of the set M' of $B_2 B_3$ doublets: Exactly as before, define the sets

$$M_0' = \{NC\}$$
$$M_1' = [(1,1) \cup (1,\beta)] \, M_0'$$
$$M_2' = (\alpha,\alpha) \, M_1' \ (\text{alternatively } M_1' = (\alpha,\alpha) \, M_2') \ ,$$

where M_1' consists of the doublets $B_2 B_3$ ending in a strong base (NS) and M_2' of the doublets ending in a weak base (NW).

Then

$$M' = M_1' \cup M_2'$$

can be expressed as

$$M' = [(1,1) \cup (1,\beta)] \, [(1,1) \cup (\alpha,\alpha)] \, M_0' \ .$$

Notice that the operator acting on M_0' has the same functional form as the operator acting on M_0 above, except that β acts as the third base instead of the first.

The sets M_1' and M_2' are invariant under the $(1,\beta)$–transformations. Then experience shows that the 32 codons in the class $N B_2 B_3$, with $B_2 B_3$ in M_1' or M_2' constitute a complete code codifying for the 20 amino acids and terminator signal (stop–codon), if allowance is made for deviating codon–assignments found in Mitochondria [17]. For the codons in M_1' this is true in the universal code; for codons in M_2' AUA should codify for M instead of I and UGA for W instead of stop signal. Both changes have been observed in Mitochondria. This more symmetric code has been considered more similar to an archetypal code than the universal code [17]. Only after the last attribute H_3 was introduced the universal code was obtained, with the split of AUR into AUA (I) and AUG (M) and UGR into UGG (W) and UGA (t).

It has been speculated that primordial genes could be included in a 0.55 kb open reading frame [18]. The same authors calculated that with two stop codons this open reading frames would have appeared too frequently. From the present view the assignment of UGA to a stop codon was a late event that optimized this frequency (this interpretation differs from the one proposed in [18,19] where a primordial code with three stop codons) is assumed. Other deviations of the universal code most likely also occurred in the last stages of the code's evolution.

In the same way as before the sixteen $B_2 B_3$ doublets can be represented as the vertices of a four–dimensional hypercube (Fig. 7). The sets M_1' and M_2' are also located in compact regions. Codons with $B_2 B_3$ in M_1' are frequently used in eukaryotes. In contrary, codons with $B_2 B_3$ in M_2' are frequently used in prokaryots. The described structure of the code allows a modulation of the codon–anticodon interaction energy [20].

V. EXAMPLES

Besides the results mentioned in the last section which refer to codon doublets, to further illustrate the significance of proposed approach, we are going to consider several examples of molecular evolution.

The first example (Fig. 2) refers to the alignment studied using the method of hierarchical analysis of residue conservation by Livingstone and Barton (Fig. 2 in [21]). In position 11 appear the following amino acids R, W, H, G, D, which according to their approach have no properties in common. In Fig. 2 this cluster of amino acids is shown. By looking at the Atlas of amino acid properties [22] we see that from the properties proposed by Grantham [23] (composition, polarity and volume) apparently the only requirement for the amino acids at this

site is to maintain a certain degree of polarity. From this observation we may conclude that most probably it is an external site. Simply by looking at such a diverse set of amino acids one can hardly realize that they have clustered codons. This clustering facilitates the occurrence of mutations that in the course of evolution were fixed, in view of the low physico–chemical requirements at the site.

As a second example (Fig. 4) let us consider site 33 of the alignment of 67 $SH2$ domains, Fig. 6 of [11]. We can see from Fig. 4 that the cluster around the codon CAC (H) explains, by one–bit changes, the amino acids R, Q, L, H, D. Furthermore, a second cluster around the codon AGC (S) explains the aminoacids R, N, S, T. Finally, a silent change from AGC (S) to UCC (S) accounts for the minor appearance of the small, neutral amino acids, A, T, P. In a similar way the variation of the hyper–variable region of immunoglobulin kappa light FR1 at position 18 can be explained (Fig. 8). The number after the amino acid symbol in Fig. 8 is the number of times the amino acid occurs in the alignment in [24].
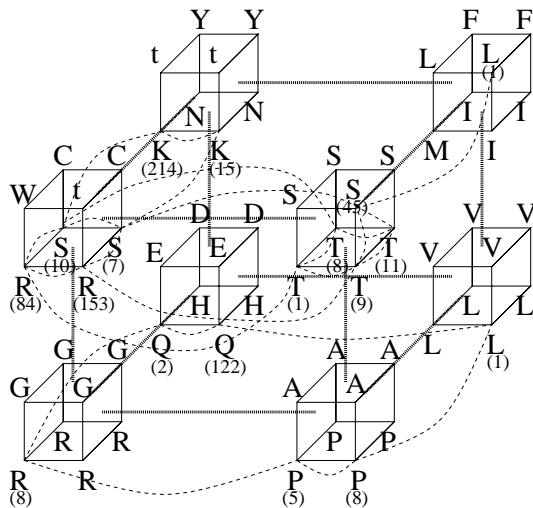

FIG. 8. The amino acid hypercube with the amino acids at position 18 of the variable region of kappa light chain displayed. The number after the amino acid symbol is the number of times the amino acid occurs in the alignment in [24].

As a third example consider the residue frequencies in 226 globins displayed in Table 3 of the paper by Bashford et al. [25]. From this table we find that there are variable positions in which one or two residues predominately occur and the rest are only marginally represented and others in which the frequencies are more evenly distributed among the amino acids. As it can be easily shown, the first class of positions may be associated, at the codon level, with one (or two) attractor node(s) and its one–bit neighbors. The second one can be associated with closed trajectories in the hypercube. The corresponding figures are not included because of lack of space.

Finally, let us discuss the "sexual PCR" experiment. In the paper by Smith [9] a table is displayed showing the positions in the TEM–1 gene where mutations occur, together with the substitutions found in the variant genes ST–1, ST–2 and ST–4 which show increased resistance to cefotaxime. We refer to the mentioned paper for further details. Locating these mutations in the hypercube (Fig. 4) one can easily convince oneself that all mutations may be accounted by one–bit changes at the codon level. Therefore only six codons (four or five aminoacids) are searched in each mutation and not 19 alternatives. This finding helps to explain why this *in vitro* realization of a "genetic algorithm" was so successful.

It is well known in the field of Genetic Algorithms that a proper encoding is crucial to the success of an algorithm. Furthermore in [26] it is shown the superiority of Gray coding over binary coding for the performance of a genetic algorithm. As it was shown above the structure of the genetic code is precisely the structure of a Gray code. Therefore it is our claim that this is one of the reasons why very efficient variants were found after very few rounds of recombination. Most probably other reasons are: the initial population was not random, but consisted of selected sequences and these sequences were very similar among themselves. This explanation of the results of Stemmer's experiment differs from the explanation advanced by Smith [9].

VI. CONCLUDING REMARKS

The present approach goes beyond the usual analyses in terms of single base changes, because it takes into account the two characters of each base and therefore it represents one–bit changes. Besides, the base position within the codon is also considered. The fact that single bit mutations occur frequently is expected from probabilistic arguments. However, one could not expect, a priori, that a cluster of mutations would correspond, at the amino acid level, to a cluster of amino acids fixed by natural selection. We have found that this situation presents itself for many positions of homologous protein sequences of many different families (results not included). The structure of the code facilitates evolution: the variation found at the variable positions of proteins do not corresponds to random jumps at the codon level, but to well defined regions of the hypercube. Finally, the Gray code structure of the genetic code helps to explain the success of "Sexual PCR" experiments.

ACKNOWLEDGMENTS

Werner Ebeling and Michael Conrad for encouraging comments.

* e–mail: jimm@udlapvms.pue.udlap.mx
+ e–mail: delamora@speedy.coacade.uv.mx
** e–mail: thorsten@hlrsun.hlrz.kfa–juelich.de
http://summa.physik.hu–berlin.de:80∼thorsten/

[1] W. P. C. Stemmer, Rapid evolution of a protein in vitro by DNA shuffling, *Nature*, **370**, 389–391 (1994).

[2] F. Crick, Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, **19** 548–555 (1966).

[3] Z. Harris, *Language and Information.* Columbia University Press (New York, 1988).

[4] M. A. Jiménez–Montaño, C. R. de la Mora–Basáñez, and T. Pöschel, On the Hypercube Structure of the genetic Code *preprint* (1994).

[5] R. W. Hamming, *Bell Syst. Tech. J.*, **29**, 147–160 (1950); W. Ebeling and R. Feistel, *Physik der Selbstorganisation und Evolution.* Akademie–Verlag (Berlin, 1982); W. Ebeling, R. Feistel, and M. A. Jiménez–Montaño, On the theory of stochastic replication and evolution of molecular sequences, *Rostocker Physikalische Manuskripte*, **2**, 105–127 (1977).

[6] L. Dillon, *The genetic mechanism and the origin of life.* Plenum Press (1978).

[7] J. Wong, A co–evolution theory of the genetic code. *Proc. Nat. Acad. Sci. USA*, **72**, 1909–1912 (1975).

[8] J. Wong, The evolution of a universal genetic code, *Proc. Nat. Acad. Sci. USA*, **73**, 2336–2340 (1976).

[9] G. P. Smith, The progeny of sexual PCR, *Nature*, **370**, 324–325 (1994).

[10] M. Eigen, Selforganization of matter and the evolution of biological macromolecules, *Naturwiss.*, **58**, 465–523 (1971).

[11] M. Guéron, E. Charretier, M. Kochoyan, and J. L. Leroy, Applications of imino proton exchange to nucleic acid kinetics and structures, in: *Frontiers of NMR in Molecular Biology.*, 225–238 (Alan R. liss Inc., 1990).

[12] R. Swanson, A unifying concept for the amino acid code, *Bull. Math. Biol.*, **46**, 187–204 (1984).

[13] M. A. Jiménez–Montaño, On the syntactic structure and redundancy distribution of the genetic code, *BioSystems*, **32**, 11–23 (1994).

[14] H. J. Danckwerts and D. Neubert, symmetries of genetic code–doublets, J. Mol. Evol., **5**, 327–332 (1975).

[15] U. Lagerkvist, "Two out of three": an alternative method for codon reading, *Proc. Nat. Acad. Sci. USA*, **75**, 1759–1762 (1978); Unorthodox codon reading and the evolution of the genetic code, *Cell*, **23**, 305–306 (1981).

[16] M. O. Bertman and J. R. Jungck, Group graph of the genetic code, *J. Heredity*, **70**, 379–384 (1979).

[17] T. H. Jukes, Evolution of the amino acid code, in: M. Nei and R. K. Koehn (eds.), *Evolution of Genes and Proteins*, Sinauer Associates Inc, 191–207 (Sunderland, Mass, 1983).

[18] H. Naora, K. Miyahara, and R. N. Curnow (1987). Origin of non-coding DNA sequences: molecular fossils of genome evolution, *Proc. Natl. Acad. Sci. USA*, **84**, 6195–6199 (1987).

[19] R. R. Brentani, Complementary hydropathy and the evolution of interacting polypeptides, *J. Mol. Evol.*, **31**, 1239–243 (1990).

[20] H. Grosjean, D. Sankoff, W. Min Jou, W. Fiers, and R. J. Cedergren, Bacteriophage MS2 RNA: A correlation between the stability of the codon: anticodon interaction and the choice of code words, *J. Mol. Evol.*, **12**, 113–119 (1978); distribution of the genetic code, *BioSystems*, **32**, 11–23.

[21] C. D. Livingstone and G. J. Barton, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation, *CABIOS*, **9**, 745–756 (1993).

[22] K. Nakai, A. Kidera, and M. Kanehisa, Cluster analysis of amino acid indices for prediction of protein structure and function, *Prot. Eng.*, **2**, 93–100 (1988).

[23] R. Grantham, Amino acid difference formula to help explain protein evolution, *Science*, **185**, 862–864 (1974).

[24] Kabat et al. *Sequences of proteins of immunological interest*, 5th ed. NIH (Bethesda, 1991).

[25] D. Bashford, C. Chothia, and A. M. Lesk, Determinants of a protein fold, *J. Mol. Biol.*, **196**, 199–216 (1987).

[26] R. A. Caruana and J. D. Schaffer, Representation and hidden bias: Gray vs. binary coding for genetic algorithms, in: J. Laird (ed.), *Proceedings of the Fifth International Conference on Machine Learning*, Morgan Kauffman Publ. Inc., 153–161 (San Mateo, 1988).